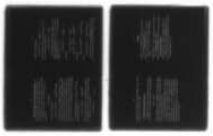
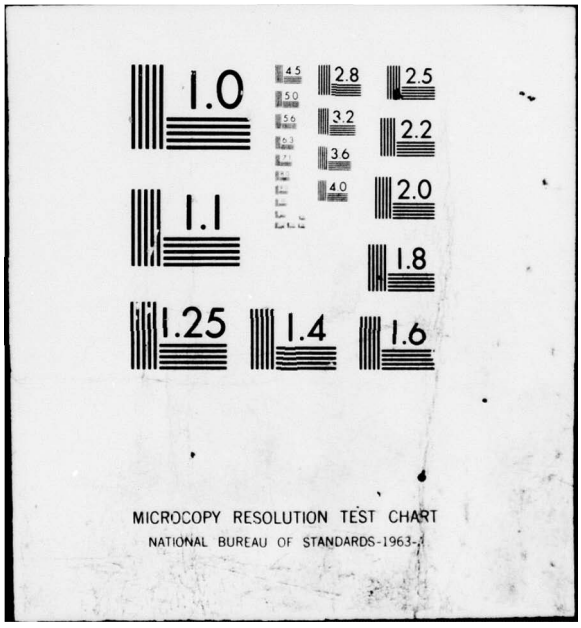
AD-A068 148 TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS F/G 12/1
DENSITY QUANTILE ESTIMATION APPROACH TO STATISTICAL DATA MODELL--ETC(U)
MAR 79 E PARZEN DAAG29-78-G-0180

UNCLASSIFIED TR-A-5 ARO-16228.5-M NL

1 OF 1
AD A068148



END
DATE FILMED
6-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

BEST AVAILABLE COPY

ARO | 16228.5-M

(18) (19)

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843

INSTITUTE OF STATISTICS
Phone 713 845-3141

AD A068148

DENSITY QUANTILE ESTIMATION APPROACH TO STATISTICAL DATA MODELLING

by Emanuel Parzen

Institute of Statistics, Texas A&M University

Technical Report No. A-5

March 1979

Texas A & M Research Foundation
Project No. 3861

"Maximum Robust Likelihood Estimation and Non-parametric Statistical Data Modeling"

Sponsored by the U.S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

DDC FILE COPY

DDC RECEIVED MAY 2 1979

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: Technical Report No. A-5	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Density Quantile Estimation Approach to Statistical Data Modelling	5. TYPE OF REPORT & PERIOD COVERED: (9) Technical rept. / 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) Emanuel Parzen	8. CONTRACT OR GRANT NUMBER(s): (15) DAAG29-78-G-0180
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas A&M University, Institute of Statistics, College Station, TX 77843	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (12) 14p.
11. CONTROLLING OFFICE NAME AND ADDRESS: Army Research Office, Research Triangle Park, NC 27709	12. REPORT DATE: Mar 1979 / 13. NUMBER OF PAGES: 26
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (14) TR-A-5	15. SECURITY CLASS. (of this report): Unclassified / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The findings of this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistical data modelling, quantile function, density-quantile function, kernel estimation, autoregressive estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper describes the density-quantile function approach to statistical analysis of a sample as involving five phases requiring the study of various population raw and smoothed quantile and density-quantile functions. The phases can be succinctly described in terms of the notation for the functions studied: (i) Q, fQ, q, (ii) $\hat{Q}$, $\hat{q}$, (iii) $\hat{fQ}$, (iv) $\hat{fQ}$, $\hat{d}$, $d(u) = f_0Q_0(u)q(u)/\sigma_0$, $\sigma_0 = \int_0^1 f_0Q_0(u)q(u)du$, (v) $\hat{Q} = \hat{\mu} + \hat{\sigma}Q_0$.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

79 04 26 031

347 380

DENSITY QUANTILE ESTIMATION APPROACH TO STATISTICAL DATA MODELLING

by

Emanuel Parzen

Institute of Statistics

Texas A&M University

Abstract

This paper describes the density-quantile function approach to statistical analysis of a sample as involving five phases requiring the study of various population raw and smoothed quantile and density-quantile functions. The phases can be succinctly described in terms of the notation for the functions studied: (i) Q, fQ, q, (ii) $\tilde{Q}$, $\tilde{q}$, (iii) $\tilde{fQ}$, (iv) $\hat{fQ}$, $\hat{d}$, $d(u) = f_0Q_0(u)q(u)/\sigma_0$, $\sigma_0 = \int_0^1 f_0Q_0(u)q(u)du$, (v) $\hat{Q} = \hat{\mu} + \hat{\sigma}Q_0$.

Research supported by grant DAAG29-78-G-0180 from the Army Research Office.

0. Introduction

The density-quantile function approach to modeling a statistical data set consisting of a sample of observations of a random variable X regards the process of statistical data analysis as involving five phases.

(1) Probability based data analysis. When asked the exploratory question "here is a data set; what can be concluded," what we desire to draw conclusions about is the probability distribution from which the sample purports to be a representative sample. Standard statistical theory is concerned with inferring from a sample the properties of a random variable that are expressed by its distribution function $F(x) = Pr[X \leq x]$ and its density function $f(x) = F'(x)$. I propose that greater insight will be obtained by formulating conclusions in terms of the qualitative and quantitative behavior of the quantile function $Q(u) = F^{-1}(u)$, $0 \leq u \leq 1$, and the density-quantile function $fQ(u) = f(Q(u))$, $0 \leq u \leq 1$. We should become familiar with the possible shapes these functions could have.

(2) Sample Quantile Function. Much current statistical theory is concerned with the properties of a sample that can be expressed in terms of its sample distribution function $\tilde{F}(x)$, $-\infty < x < \infty$, defined by

$$\tilde{F}(x) = \text{proportion of the sample with values} \leq x .$$

I propose that the basic descriptive statistics of a sample is its sample quantile function $\tilde{Q}(u)$ defined so that it has a derivative $\tilde{q}(u) = \tilde{Q}'(u)$, called the sample quantile-density function. Exploring the data for

patterns, as well as modelling the data, consists of examining how well various theoretical quantile functions Q(u) match, or fit, $\tilde{Q}$.

(3) Sample Density-Quantile Function $\tilde{fQ}$. The most widely used graphical procedure for inspecting a sample is the histogram. I propose as an alternative a raw estimator $\tilde{fQ}$ of fQ, which can be obtained in several ways. The graph of $\tilde{fQ}$ provides insights into the type of distribution which the data possesses, including the following types: Symmetric; J-shaped; Skewed to the right; Skewed to the left; Uniform; Normal; Exponential; Short-tailed (limited type); Long-tailed (Cauchy or Student t type); Exponential-tailed (Weibull type); Biomodal (or multimodal); Zeroes in density; Outliers; Discrete (infinities in density).

(4) Smoothed Density-Quantile Function $\hat{fQ}$. The qualitative behavior or shape of the density-quantile function fQ classifies the type of probability distribution which a random variable possesses. To answer estimation questions about a random variable, we need quantitative estimators $\hat{fQ}$ of fQ which can be accomplished by a variety of smoothing or density estimation methods. I propose that an easily implementable procedure is provided by autoregressive smoothing of $f_0Q_0(u)\tilde{q}(u)$, where $f_0Q_0(u)$ is a "flattening" function specified by the statistician. This approach also provides goodness-of-fit tests of the hypothesis

$$H_0: F(x) = F_0\left(\frac{x-\mu}{\sigma}\right) , \quad Q(u) = \mu + \sigma Q_0(u)$$

where F_0 is a specified distribution function with quantile function $Q_0(u)$, and μ and σ are location and scale parameters to be estimated.

(5) Parametrically Smoothed Quantile Function $\hat{Q}$. To complete the process of modeling the sample quantile function $\tilde{Q}$ (that is, fitting

$\tilde{Q}$ by a theoretical quantile function Q), one postulates a parametric model such as $Q(u) = \mu + \sigma Q_0(u)$ to be fitted to $\tilde{Q}(u)$. Parameters such as μ and σ can be efficiently estimated by regression analysis of the continuous parameter "time series" $\tilde{Q}(u)$ using the theorem that the asymptotic distribution of $fQ(u)\{\tilde{Q}(u) - Q(u)\}$ is a Brownian bridge stochastic process. A final check of the model is provided by the goodness of fit of $\hat{Q}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$ to $\tilde{Q}(u)$.

Some of the details involved in carrying out the foregoing phases of statistical data modeling are described in this paper.

1. Quantile Functions and Density Quantile Functions

Corresponding to a distribution function $F(x)$, $-\infty < x < \infty$, we define its quantile function $Q(u)$, $0 \leq u \leq 1$, to be its inverse:

$$Q(u) = F^{-1}(u) = \inf\{x: F(x) \geq u\} . \tag{1}$$

Note that we use u to denote the argument of Q; it is a variable in the unit interval: $0 \leq u \leq 1$.

Two identities are so useful that I give them names:

Correspondence Identity: $F(x) \geq u$ if, and only if, $Q(u) \leq x$; (2)

Inverse Identity: $FQ(u) = u$ if F is continuous . (3)

Three important functions are defined by

Quantile density function $q(u) = Q'(u)$ (4)

Density-quantile function $fQ(u) = f(Q(u))$ (5)

Score function $J(u) = -(fQ)'(u)$. (6)

79 04 26 031

The shapes of these functions turn out to be independent of location and scale parameters.

By differentiating the Inverse Identity, we obtain

$$\text{Reciprocal Identity:}\quad fQ(u)q(u) = 1 . \tag{7}$$

In words, the quantile-density and density-quantile function are reciprocals of each other.

One important consequence of the Reciprocal Identity is the agreement of our definition of the score function J(u) (in terms of the derivative of fQ(u)) with the definition given in the theory of non-parametric statistical inferences

$$J(u) = \frac{-f'(F^{-1}(u))}{f(F^{-1}(u))} . \tag{8}$$

It seems easier to estimate J(u) using formula (6) rather than formula (8).

In the density-quantile function approach a basic role is played by the density-quantile function of the normal distribution

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)dy , \quad \phi(y) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2} y^2) .$$

The quantile function $\Phi^{-1}(u)$ has to be computed numerically. Then one computes the density-quantile function by

$$\phi\Phi^{-1}(u) = \frac{1}{\sqrt{2\pi}} \exp -\tfrac{1}{2}\{\Phi^{-1}(u)\}^2 .$$

The score function $J(u) = \Phi^{-1}(u)$.

The exponential distribution with

$$f(x) = e^{-x} , \quad 1 - F(x) = e^{-x}$$

has quantile function

$$Q(u) = \log \frac{1}{1-u}$$

since $u = F(x)$, $x = Q(u)$ implies $1 - u = e^{-Q(u)}$; $fQ(u) = 1 - u$.

To illustrate how one obtains distribution functions from quantile functions, consider

$$Q(u) = \tfrac{1}{2}(\log \frac{1}{1-u})^{\frac{1}{2}} .$$

Consequently, $x = Q(u)$ implies $u = F(x) = 1 - e^{-4x^2}$, and $f(x) = 8x\, e^{-4x^2}$ which is the Rayleigh distribution; $fQ(u) = 4(1 - u)(-\log(1 - u))^{\frac{1}{2}}$.

An important consequence of the Correspondence and Inverse Identities are the following facts. Let U denote a random variable uniformly distributed on the interval [0, 1], and $\sim$ denote "identically distributed as." Then one can represent X by the

$$\text{Representation Identity:}\quad X \sim Q(U) . \tag{9}$$

When F is continuous

$$\text{Probability Integral Transformation:}\quad F(X) \sim U . \tag{10}$$

The representation identity plays a central role in theorem-proving because it enables one to first prove theorems for uniformly distributed random variables, and then extend to arbitrary random variables by using the representation (9) and the analytic properties of Q. This technique was first used by Scheffé and Tukey (1945). Tukey (1965) calls Q the representing function and q the sparsity function.

To simulate a sample $X_1, \ldots, X_n$ of a random variable X, one could simulate a sample $U_1, \ldots, U_n$ of U, and form $X_1 = Q(U_1), \ldots, X_n = Q(U_n)$. I understand from Professor Jim Thompson of Rice that the numerical analysis techniques are now available to make this a practical universal approach to simulating an arbitrary continuous distribution.

One may want to transform X to another random variable Y which has a specified continuous distribution function $F_0(x)$ and quantile function $Q_0(u)$. This can be accomplished using the facts

Transformation Identities: $X \sim QF_0(Y)$, $Y \sim Q_0F(X)$. (17)

Quantile Function of a Monotone Function. An extremely useful property of the quantile function is how easily it can be found for a random variable Y which is obtained from X by a monotone transformation g: $Y = g(X)$. When g is an increasing function,

$$Q_Y(u) = g(Q_X(u)) . \tag{12}$$

When g is a decreasing function,

$$Q_Y(u) = g(Q_X(1 - u)) . \tag{13}$$

Applications of these formulas are

$$Y = \mu + \sigma X , \qquad Q_Y(u) = \mu + \sigma Q_X(u) , \tag{14}$$

$$Y = -\log X , \qquad Q_Y(u) = -\log Q_X(1 - u) , \tag{15}$$

$$Y = 1/X , \qquad Q_Y(u) = 1/Q_X(1 - u) . \tag{16}$$

Moments: Moments of X are easily expressed in terms of the quantile function since

Moment Identity: $E[g(X)] = E[g(Q(u))] = \int_0^1 g(Q(u))du$. (17)

The mean is

$$\mu = \int_0^1 Q(u)du . \tag{18}$$

The median is Q(0.5). Parameters whose estimation is considered more robust are the trimean $\frac{1}{4}Q(0.25) + \frac{1}{2}Q(0.5) + \frac{1}{4}Q(0.75)$, the trimmed mean

$$\mu_p = \int_p^{1-p} Q(u)du , \tag{19}$$

and the weighted mean $\mu_w = \int_0^1 w(u)Q(u)du$, for a specified weight function w(u). Corresponding measures of variation would be functionals of the deviations $|Q(u) - \mu_w|$ of the quantile function from the representative value μ_w. The variance can be expressed

$$\sigma^2 = \int_0^1 \{Q(u) - \mu\}^2 du . \tag{20}$$

Tail behavior of probability laws. To describe all possible continuous probability distributions, it suffices to describe all possible continuous monotone functions Q(u), $0 \leq u \leq 1$. To describe all possible types of tail behavior of probability laws, it suffices to describe the behavior of Q(u) as u tends to 0 or 1. We choose to express this behavior in terms of fQ(u). Let α be a parameter satisfying $-\infty < \alpha < \infty$. We call α

(i) lower tail exponent if

$$fQ(u) \sim u^{\alpha} \quad \text{as} \quad u \to 0 ,$$

$$\alpha = \lim_{u \to 0} \frac{uJ(u)}{fQ(u)}$$

(ii) upper tail exponent if

$$fQ(u) \sim (1-u)^{\alpha} \quad \text{as} \quad u \to 1 ,$$

$$\alpha = \lim_{u \to 1} \frac{(1-u)J(u)}{fQ(u)} .$$

The parameter ranges (i) $0 \leq \alpha < 1$, (ii) $\alpha = 1$, and (iii) $\alpha > 1$ correspond to the three types of tail behavior

(i) short tails or limited type,

(ii) medium tails or exponential type,

(iii) long tails or Cauchy type.

The parameter range $\alpha < 0$ could be called super-short tails; the densities are unbounded and the corresponding characteristic functions $\varphi(t) = E[e^{itx}]$, $-\infty < t < \infty$, decay very slowly as $t \to \infty$ and are not even integrable. The general treatment of such random variables require further study. (An example is $X = \cos \pi U$).

Extreme value distributions are those corresponding to the random variables,

$$\text{(i) } \xi^{\beta} , \quad \text{(ii) } \log \xi , \quad \text{(iii) } \xi^{\beta}$$

where ξ is exponential with mean 1, and β depends on the value of α, $\beta = 1 - \alpha$. The quantile functions of Weibull distributions are

$$Q(u) = (\log \frac{1}{1-u})^{\beta} , \quad \beta = 1 - \alpha \quad \text{where} \quad 0 \leq \alpha < 1 ;$$

the extreme value distribution has quantile function

$$Q(u) = \log \log \frac{1}{1-u} , \quad \alpha = 1 .$$

Note that for $\beta = 1$, corresponding to $\alpha = 0$, the distribution is exponential.

Conditional means and an approach to Empirical Bayes Estimation.

When the distribution of observations X depend on an unknown parameter θ to be estimated, a Bayesian estimator of θ is the conditional mean $E[\theta|X]$ usually written

$$E[\theta|X] = \frac{\int_{-\infty}^{\infty} \theta f(X|\theta) g(\theta) d\theta}{\int_{-\infty}^{\infty} f(X|\theta) g(\theta) d\theta}$$

where $f(X|\theta)$ is the conditional density of X given θ, and $g(\theta)$ is the prior density of θ. Let $Q_{\theta}(u)$ denote the prior quantile function of θ;

$$Q_{\theta}(u) = G^{-1}(u) , \quad G(x) = \int_{-\infty}^{x} g(\theta) d\theta .$$

One can show that

Conditional Mean Identity:

$$E[\theta|X] = \frac{\int_0^1 Q_{\theta}(u) f(X|Q_{\theta}(u)) du}{\int_0^1 f(X|Q_{\theta}(u)) du} .$$

In practice, one may be willing to assume $f(X|\theta)$ but the prior distribution of θ is unknown. The empirical Bayes attitude is to estimate the distribution of θ from previous estimators $\hat{\theta}_1, \ldots, \hat{\theta}_n$. What our new formula for $E[\theta|X]$ indicates is that it suffices to estimate the prior quantile function $Q_{\theta}(u)$.

The conditional distribution of θ given the observations X can be evaluated using the formula, for $0 \leq p \leq 1$

$$P[\theta \leq Q_{\theta}(p)|X] = \frac{\int_0^p f(X|Q_{\theta}(u)) du}{\int_0^1 f(X|Q_{\theta}(u)) du} \tag{21}$$

To obtain a formula for the conditional quantile function of θ given X, denoted $Q_{\theta|X}(u)$, denote the right hand side of (21) by $D_{\theta|X}(p)$:

$$D_{\theta|X}(p) = \frac{\int_0^p f(X|Q_\theta(u))du}{\int_0^1 f(X|Q_\theta(u))du}, \quad 0 \le p \le 1 . \tag{22}$$

Then (21) can be written

$$F_{X|\theta}(Q_\theta(p)) = D_{\theta|X}(p) . \tag{23}$$

Then $F_{X|\theta}(x) = u$ for $x = Q_\theta(p)$ where p satisfies $D_{\theta|X}(p) = u$ whence $p = D^{-1}_{\theta|X}(u)$ and

Conditional Quantile Identity: $$Q_{\theta|X}(u) = Q_\theta(D^{-1}_{\theta|X}(u)) \tag{24}$$

which is an extremely important formula.

We might regard the conditional median $Q_{\theta|X}(0.5)$ as an estimator of θ given X; (24) says that it equals the prior quantile function Q_θ evaluated at $D^{-1}_{\theta|X}(0.5)$.

One may be able to quickly obtain insight into whether a new observation X implies a "significantly different" estimator of θ. Speaking extremely intuitively, we can form an acceptance region for the null hypothesis H_0 that the "true" value of θ is approximately the prior median $Q_\theta(0.5)$ at a level of significance α; define $p = D^{-1}_{\theta|X}(0.5)$ and call it the p-median of the observation X. Accept H_0 if p satisfies an inequality of the form $\alpha/2 \le p \le 1 - (\alpha/2)$, $p \le 1 - \alpha$, $p \ge \alpha$ (depending on whether the test is two-sided or one-sided).

Other consequences of "thinking quantile". A $(1 - \alpha)$-confidence level "prediction" interval for the values of a random variable X with a symmetric distribution could be the interval $Q(\alpha/2) \le X \le Q(1 - (\alpha/2))$. For a normal random variable $Q(u) = \mu + \sigma\Phi^{-1}(u)$ and the prediction interval is $|X - \mu| \le \sigma\Phi^{-1}(\alpha/2)$. If X is an unbiased estimator $\hat{\theta}$ of a

parameter θ, the confidence interval is $|\hat{\theta} - \theta| \le \sigma\Phi^{-1}(\alpha/2)$ where σ is the standard deviation of $\hat{\theta}$. Many statistics text books use the intuitive notation $z(\alpha/2)$ for the mathematically precise $\Phi^{-1}(\alpha/2)$.

It should be noted that the shape of fQ and q is independent of location and scale parameters in the sense that the following equations are equivalent

$$F(x) = F_0(\tfrac{x-\mu}{\sigma}) , \quad Q(u) = \mu + \sigma Q_0(u) ,$$

$$f(x) = \tfrac{1}{\sigma} f_0(\tfrac{x-\mu}{\sigma}) , \quad fQ(u) = \tfrac{1}{\sigma} f_0 Q_0(u) .$$

A symmetric density

$$f(x) = f(-x)$$

is equivalent to

$$Q(1 - u) = -Q(u) , \quad fQ(1 - u) = fQ(u) .$$

2. Sample Quantile Function

The basic definition of the sample quantile function $\tilde{Q}(u)$, $0 \le u \le 1$, is

$$\tilde{Q}(u) = \tilde{F}^{-1}(u) = \inf\{x: \tilde{F}(x) \ge u\}$$

where $\tilde{F}(x)$ is the sample distribution function. Its probability theory is explored in elegant detail in recent work of Czorgo and Revesz (1978) who show that

$$fQ(u)\{Q(u) - \tilde{Q}(u)\} \sim B(u) , \text{ a Brownian Bridge process .}$$

For data analysis we prefer a definition of $\tilde{Q}(u)$ which is differentiable so that a sample quantile-density function

$$\tilde{q}(u) = \tilde{Q}'(u)$$

can be defined. Its basic property is that

$$\tilde{q}(u) \sim \text{exponential with mean } q(u) .$$

In this section we emphasize the computational details involved in computing $\tilde{Q}(u)$ under various definitions. An alternative name and notation for the sample quantile function $\tilde{Q}(u)$, $0 \leq u \leq 1$ is the percentile function $X(p)$, $0 \leq p \leq 1$. It can be given a variety of slightly different definitions, which are all equivalent in the limit as the sample size tends to ∞. Intuitively, $X(p)$ is a value such that $100p\%$ of the numbers in the sample are less than or equal to $X(p)$, and $100(1 - p)\%$ of the numbers in the sample are greater than or equal to $X(p)$. Mathematically one might define $X(p)$ as the inverse of the sample distribution function, denoted by $\tilde{F}(x)$ or $P(x)$:

$$X(p) = P^{-1}(p) = \inf\{x: P(x) \geq p\} , \qquad 0 \leq p \leq 1 .$$

When P is continuous, $P(X(p)) = p$. When $P(x)$ is also strictly increasing, $x = X(p)$ if and only if $p = P(x)$.

When the sample of size n is available as a set of numbers (the case of a histogram is considered next), one can form the order statistics of the sample, denoted

$$X_{1;n} \leq X_{2;n} \leq \cdots \leq X_{n;n} ;$$

the order statistics are the numbers in the sample arranged in non-decreasing order. To define $X(p)$ we associate with p an index $i(p)$; possible definitions for $i(p)$ are

$$i(p) = np + \tfrac{1}{2} \quad \text{or} \quad i(p) = (n + 1)p .$$

We consider explicitly only the first definition of $i(p)$. When $i(p)$ is an integer, we define

$$X(p) = X_{i(p);n} .$$

This is equivalent to defining the value of $X(p)$ for $p = (2j - 1)/2n$ to be $X_{j;n}$. For other values of p we define $X(p)$ to be either piecewise constant or by linear interpolation. The piecewise constant definition is most convenient for hand calculation; it is equivalent to: if $i(p)$ is not an integer, choose an integer j so that $j < i(p) < j + 1$, and define $X(p) = \tfrac{1}{2}\{X_{j;n} + X_{(j+1);n}\}$. The linear interpolation definition is preferred in general because its derivative exists.

We argue that all the statistical information in a sample is contained in (and indeed is easily extracted from) the percentile function. Examples of how easily statistical measures are expressed in terms of percentile functions are:

median	$X(0.5)$
quartiles	$X(0.75, X(0.25)$
mid-quartile	$\tfrac{1}{2}\{X(0.25) + X(0.75)\}$
tri-mean	$\tfrac{1}{4}X(0.25 + \tfrac{1}{2}X(0.5) + \tfrac{1}{4}X(0.75)$
mean	$\bar{X} = \int_0^1 X(p)dp$

In words, the mean of the sample is the mean of the percentile function.

p-trimmed mean $\frac{1}{1-2p}\int_{-p}^{p} X(u)du$

p-Winsorized mean $p\,X(p) + \int_{p}^{1-p} X(u)du + p\,X(1-p)$

Interquartile range $X(0.75) - X(0.25)$

Standard deviation $SD = \sqrt{\text{Variance}}$

Variance $\int_0^1 (X(p) - \bar{X})^2\,dp$

In words, the standard deviation is the root mean square of the deviations of the percentile function from the representative value $\bar{X}$.

Wilcoxon test statistic $\int_0^1 p\ \text{sign}\ X(p)\,dp$

Sign test statistic $\int_0^1 \text{sign}\ X(p)\,dp$

Histograms. When a sample of size n is described by a histogram, one defines k intervals by endpoints x_{j-1}, x_j for $j = 1, \ldots, k$. The number of observations with values in $x_{j-1} < x \le x_j$ is called the frequency of the interval. For $j = 1, \ldots, k$, define

$$p_j = \frac{n_j}{n}\ ,\quad f_j = \frac{p_j}{x_j - x_{j-1}}$$

called respectively the relative frequency and the density in the j-th interval. To describe a sample, one gives the table:

Interval	Frequency	Relative Frequency	Cumulative Relative Frequency	Density
x_{j-1} to x_j	n_j	p_j	$p_1 + p_2 + \ldots + p_j$	f_j

and plots the histogram function h(x) defined by

$$h(x) = f_j\ , \qquad x_{j-1} < x < x_j$$
$$= \tfrac{1}{2}(f_j + f_{j+1})\ , \quad x = x_j\ .$$

The sample distribution function F(x) is defined to be integral of the histogram:

$$\tilde{F}(x) = \int_{-\infty}^{x} h(y)dy\ .$$

$\tilde{F}(x)$ is computed as follows: for $j = 1, 2, \ldots, k$

$$\tilde{F}(x_j) = p_1 + p_2 + \ldots + p_j\ .$$

At other values of x, define $\tilde{F}(x)$ by linear interpolation. Thus for $x_{j-1} \le x \le x_j$

$$\tilde{F}(x) = \tilde{F}(x_j)\,\frac{(x-x_{j-1})}{(x_j-x_{j-1})} + \tilde{F}(x_{j-1})\,\frac{(x_j-x)}{(x_j-x_{j-1})}\ .$$

Given an histogram table the percentile or sample quantile function is defined by

$$\tilde{Q}(0) = x_0\ ,$$

$$\tilde{Q}(u) = x_j \quad \text{for} \quad u = p_1 + p_2 + \ldots + p_j$$

and by linear interpolation for other values of u in $0 < u < 1$.

It is to be emphasized that computing and plotting the sample quantile function is the first step in statistical data analysis, expecially when combined with the box-plot technique of Tukey (1977) to provide Quantile-Box Plots (Parzen (1978)).

3. Raw Density-Quantile Function $\tilde{fQ}$.

Having computed and plotted $\tilde{Q}$, the next step in statistical data analysis is to compute a raw estimator of the density-quantile function which we denote by $\tilde{fQ}(u)$.

When the data is reported as a histogram, the raw density quantile function, denoted $\tilde{fQ}(u)$, is defined by

$$\tilde{fQ}(u) = h(\tilde{Q}(u)) ;$$

it satisfies for u in $p_1 + \ldots + p_{j-1} < u < p_1 + \ldots + p_j$

$$\tilde{fQ}(u) = f_j .$$

For $u = p_1 + \ldots + p_j$, $\tilde{fQ}(u) = \frac{1}{2}\{f_j + f_{j+1}\}$.

When the data is reported as $\tilde{Q}$, a raw density-quantile function is formed from a raw estimator $q^*(u)$ of the quantile-density function as follows:

$$\tilde{fQ}(u) = 1/q^*(u) .$$

The basic requirement for $q^*(u)$ is that it be slightly smoother than $\tilde{q}(u)$. In general, to form a smooth estimator $q^*(u)$ of $q(u)$ one could use a kernel estimator

$$q^*(u) = \int_0^1 \tilde{q}(p) \frac{1}{h} K(\frac{u-p}{h})dp$$

for a suitable kernel K and band width h.

At this stage we only seek to smooth $\tilde{q}$ enough so that it would be statistically meaningful to form its reciprocal. Therefore, we recommend computing q^* at equi-spaced values $u = h, 2h, \ldots, 1 - 2h, 1 - h$ by

$$q^*(jh) = \{\tilde{Q}((j + 1)h) - \tilde{Q}((j - 1)h)\} \div 2h .$$

Define $q^*(u)$ for other values of u by linear interpolation.

The properties of $q^*(u)$ are given by Bofinger (1975) who shows that it is asymptotically normal with asymptotic variance and mean

$$\mathrm{Var}[q^*(u)] = \frac{1}{2nh} q^2(u)$$

$$E[q^*(u)] = q(u) + \frac{1}{6} h^2 q''(u)$$

$$\mathrm{Bias}[q^*(u)) = \frac{1}{6} h^2 q''(u) .$$

Let h_{min} denote the choice of h which minimizes mean square error,

$$\text{Mean Square Error} = \text{Variance} + \text{Bias Squared} ;$$

one can show that one should choose h_{min} so that

$$\text{Variance} = 4 \text{ Bias Squared}$$

whence

$$\frac{1}{2nh_{min}} (q(u))^2 = \frac{1}{9} h_{min}^4 (q''(u))^2 .$$

The following important conclusion has been proved:

$$h_{min} = (\frac{1}{n})^{1/5} C(u)$$

where

$$C(u) = (4.5)^{1/5} (\frac{q(u)}{q''(u)})^{2/5} .$$

We seek a lower bound for h_{min} to yield reasonably accurate estimators. One can argue that a "worst" case is the Cauchy distribution for which $fQ(u) = (\sin \pi u)^2/\pi$, and C(u) has values given by the following table (taken from Bofinger (1975)):

u	0.50	0.60	0.70	0.80	0.90	0.95
C(u)	.41	.37	.28	.19	.11	.06

For $n = 2^5 = 32$, $h_{min} = \frac{1}{2}C(u)$; for $n = 4^5 = 1024$, $h_{min} = \frac{1}{5}C(u)$.

What we would like to do in practice is to compute $\tilde{f}Q(u)$ at an equi-spaced grid of values h, 2h, 3h, ..., $1 = 2h$, $1 - h$. A choice of $h = 0.05$ or 0.025 yields the amount of smoothing that is reasonable for the worst case (long tailed densities). The optimal choice of h would undoubtedly be larger (especially for values of u near 0.5). The path we follow to obtain an optimally smoothed estimator is to use preflattened smoothing, defined in the next section.

4. Smoothed Density-Quantile Function $\hat{f}Q$.

One can develop many approaches to forming smooth functions $\hat{q}(u)$ and $\hat{f}Q(u)$ which can be regarded as estimators of $q(u)$ and $fQ(u)$. The approach we recommend has three important features: (1) it smooths a pre-flattened sample quantile-density, (2) it uses autoregressive smoothers, and (3) it provides goodness-of-fit tests for hypotheses that the true distribution function belongs to a specified location and scale parameter family of distribution functions.

Goodness of Fit Tests. Goodness of fit tests are concerned with testing hypotheses about the distribution function $F(x)$ of a random variable X. Let $F_0(x)$ be a specified distribution function with quantile function $Q_0(u)$, and density- quantile function $f_0Q_0(u)$. The unrealistic case of a simple hypothesis

$$H_0: F(x) = F_0(x) , \quad Q(u) = Q_0(u)$$

is considered first to illustrate how one formulates goodness of fit tests. Conventional statistics tests recommend transforming X to $U = F_0(X)$, and testing whether U is uniform, using texts based on the sample distribution function of U.

Our first departure from the conventional approach is to emphasize using tests based on the sample quantile function of U, which we denote by $\tilde{D}(u)$. One can express it in terms of the sample quantile function $\tilde{Q}(u)$ of X by

$$\tilde{D}(u) = F_0(\tilde{Q}(u)) .$$

A more realistic hypothesis for $F(x)$ is a location-scale parameter model,

$$H_0: F(x) = F_0(\tfrac{x-\mu}{\sigma}) , \quad Q(u) = \mu + \sigma Q_0(u) .$$

Let $\hat{\mu}$ and $\hat{\sigma}$ be estimators of the unknown parameters. Conventional texts recommend forming

$$\hat{U}_1 = F_0(\tfrac{X_1-\hat{\mu}}{\hat{\sigma}}), \ \ldots, \ \hat{U}_n = F_0(\tfrac{X_n-\hat{\mu}}{\hat{\sigma}}) ,$$

and using tests based on their sample distribution function. We would prefer tests based on the sample quantile function which we now denote $\tilde{D}_0(u)$; it can be expressed:

$$\tilde{D}_0(u) = F_0(\tfrac{\tilde{Q}(u)-\hat{\mu}}{\hat{\sigma}}) .$$

A method of generalizing this procedure, which avoids estimating μ and σ, is suggested by forming the density

$$\tilde{d}_0(u) = \tilde{D}_0'(u)$$
$$= f_0(\tfrac{\tilde{Q}(u)-\hat{\mu}}{\hat{\sigma}})\tilde{q}(u)\ \tfrac{1}{\hat{\sigma}}$$

where $\tilde{q}(u) = \tilde{Q}'(u)$ is the sample quantile-density function.

An important formula for $\tilde{q}(u)$ is: for $j = 1, 2, \ldots, n-1$

$$\tilde{q}(u) = n(X_{j;n} - X_{j-1;n}) , \quad \frac{2j-1}{2n} < u < \frac{2j+1}{2n} ;$$

the values of $\tilde{q}(u)$ are called spacings. It should be noted that the statistical properties of $\tilde{q}(u)$ are isomorphic to those of the sample spectral density of a stationary time series.

A new approach is to define a new density function

$$\tilde{d}(u) = f_0 Q_0(u) \tilde{q}(u) \frac{1}{\tilde{\sigma}_0}$$

where

$$\tilde{\sigma}_0 = \int_0^1 f_0 Q_0(u) \tilde{q}(u) du .$$

Note that if $f_0 \tilde{Q}_0(u) Q(u) = 0$ at $u = 0$ and 1, one can write

$$\tilde{\sigma}_0 = \int_0^1 J_0(u) \tilde{Q}(u) du$$

which is a scale estimator that often coincides with the usual estimator when H_0 holds.

We call $\tilde{d}(u)$ the weighted spacings function,

$$\tilde{D}(u) = \int_0^u \tilde{d}(t) dt$$

the cumulative weighted spacings function, and

$$\tilde{\phi}(v) = \int_0^1 e^{2\pi i u v} \tilde{d}(u) du , \quad v = 0, \pm 1, \ldots$$

the pseudo-correlations.

We can regard $\tilde{d}(u)$ as an "estimator" (unfortunately, only consistent when used as the integrand of an integral) of

$$d(u) = f_0 Q_0(u) q(u) \frac{1}{\sigma_0}$$

where

$$\sigma_0 = \int_0^1 f_0 Q_0(u) q(u) du .$$

Under the null hypothesis, $d(u)$ is identically 1.

Parzen (1979) introduces autoregressive estimators $\hat{d}(u)$ of $d(u)$ which can estimate $\hat{d}(u) = 1$ a specified proportion of the time when H_0 is true. One thus simultaneously tests whether H_0 is true, and estimates $d(u)$ when H_0 is rejected.

Autoregressive Estimation of the Density Quantile Function. To obtain an estimator $\hat{fQ}(u)$ of $fQ(u)$ which has good mean square error properties at each point u, and is not too wiggly as a function of u, it is desirable to use a parametric representation of $fQ(u)$ to estimate it. The hypothesis H_0: $Q(u) = \mu + \sigma Q_0(u)$ is equivalent to the representation

$$fQ(u) = \frac{1}{\sigma} f_0Q_0(u) .$$

A more general representation is

$$fQ(u) = C_m|1 + \alpha_m(1)e^{2\pi iu} + \ldots + \alpha_m(m)e^{2\pi ium}|^2 f_0Q_0(u)$$

for some integer m, $C_m > 0$, and complex coefficients $\alpha_m(1), \ldots, \alpha_m(m)$. The "base" function $f_0Q_0(u)$ can often be suggested by the data through an inspection of $\tilde{fQ}(u)$. One would like to choose $f_0Q_0(u)$ so as to reduce the number m of parameters in the representation. One can show that under rather general conditions to any specified f_0Q_0 there exists (in the limit as m tends to ∞) a representation for fQ of the foregoing form.

The foregoing representation for fQ implies that d(u) has the representation (for some $K_m > 0$)

$$d(u) = K_m|1 + \alpha_m(1)e^{2\pi iu} + \ldots + \alpha_m(m)e^{2\pi ium}|^{-2} .$$

In words, d(u) is the reciprocal of the square modulus of a polynomial. Such a representation may appear at first sight as unpromising. However it is equivalent to the Fourier transform

$$\phi(v) = \int_0^1 e^{2\pi iuv} d(u)du , \qquad v = 0, \pm 1, \ldots$$

satisfying a difference equation

$$\phi(v) + \alpha_m(1)\phi(1 - v) + \ldots + \alpha_m(m)\phi(m - v) = 0 , \qquad v > 0 ,$$

which can be used to determine the coefficients $\alpha_m(j)$ if $\phi(v)$ are known. Further, one can determine K_m by

$$K_m = \phi(0) + \alpha_m(1)\phi(1) + \ldots + \alpha_m(m)\phi(m) .$$

One can form a sequence $\hat{fQ}_m(u)$ of smooth estimators of $fQ(u)$ as follows. First form estimators $\tilde{\phi}(v)$ of $\phi(v)$. Second, for each m, determine coefficients $\hat{\alpha}_m(j)$, $j = 1, \ldots, m$, by solving the system of equations, with $v = 1, \ldots, m$,

$$\tilde{\phi}(v) + \hat{\alpha}_m(1)\tilde{\phi}(1 - v) + \ldots + \hat{\alpha}_m(m)\tilde{\phi}(m - v) = 0 .$$

Third, define

$$\hat{K}_m = \tilde{\phi}(0) + \hat{\alpha}_m(1)\tilde{\phi}(1) + \ldots + \hat{\alpha}_m(m)\tilde{\phi}(m)$$

$$\hat{d}_m(u) = K_m|1 + \hat{\alpha}_m(1)e^{2\pi iu} + \ldots + \hat{\alpha}_m(m)e^{2\pi ium}|^{-2}$$

Fourth, define

$$\hat{fQ}_m(u) = C_m|1 + \hat{\alpha}_m(1)e^{2\pi iu} + \ldots + \alpha_m(m)e^{2\pi imu}|^2 f_0Q_0(u)$$

where

$$C_m^{-1} = \int_0^1 |1 + \hat{\alpha}_m(1)e^{2\pi iu} + \ldots + \alpha_m(m)e^{2\pi ium}|^2 f_0Q_0(u)du .$$

The crucial question is the order determination problem; find a value of the order m, to be denoted $\hat{m}$, such that $\hat{d}_{\hat{m}}(u)$ is an "optimal"

estimator of $d(u)$ and $\hat{fQ}_m(u)$ is an "optimal" estimator of $fQ(u)$. Further research needs to be done on this problem.

5. Parametric Smoothed Quantile Functions $\hat{Q}$.

Estimation of the fQ function only determines Q up to location and scale parameters. Thus the parametric model for a true quantile function

$$Q(u) = \mu + \sigma Q_0(u)$$

where Q_0 is known, and μ and σ are parameters to be estimated, can arise either from theory or as part of the process of fitting a smooth quantile function to a empirical quantile function $\tilde{Q}$.

Parzen (1979) discusses efficient estimation of the location and scale parameters μ and σ in the parametric model $Q(u) = \mu + \sigma Q_0(u)$ for the true quantile function Q. Equivalent to using a restricted set of order statistics $X_{np;n}, \ldots, X_{nq;n}$ (or a trimmed sample) is using the sample quantile function $\tilde{Q}(u)$, $p \leq u \leq q$. One can form asymptotically efficient estimators denoted $\mu_{p,q}$ and $\sigma_{p,q}$, using normal equations in suitable linear functionals in $\tilde{Q}$.

References

Bofinger, E. (1975), "Estimation of a density function using order statistics," Austral. J. Statistics 17, 1-7.

Bofinger, E. (1975), "Non-parametric estimation of density for regularly varying distributions," Austral. J. Statistics 17, 192-195.

Czorgo, M. and Revesz, P. (1978), "Strong Approximations of the Quantile Process," Annals Statistics 6, 882-897.

Parzen, E. (1979), "Non-parametric Statistical Data Modeling," Journal American Statistical Association.

Scheffé, H. and Tukey, J. W. (1945), "Non-parametric estimation, I Validation of order statistics," Ann. Math. Statist. 16, 187-192.

Tukey, J. N. (1965), "Which part of the sample contains the information," Proc. Nat. Acad. Sci. 53, 127-134.